Informed Sound Source Localization Using Relative Transfer Functions for Hearing Aid Applications

Farmani, Mojtaba; Pedersen, Michael Syskind; Tan, Zheng-Hua; Jensen, Jesper

Published in:
IEEE/ACM Transactions on Audio, Speech, and Language Processing

DOI (link to publication from Publisher):
10.1109/TASLP.2017.2651373

Publication date:
2017

Document Version
Accepted author manuscript, peer reviewed version

Informed Sound Source Localization Using Relative Transfer Functions for Hearing Aid Applications

Mojtaba Farmani, Michael Syskind Pedersen, Zheng-Hua Tan, and Jesper Jensen,

***Abstract*—Recent hearing aid systems (HASs) can connect to a wireless microphone worn by the talker of interest. This feature gives the HASs access to a noise-free version of the target signal. In this paper, we address the problem of estimating the target sound direction of arrival (DoA) for a binaural HAS given access to the noise-free content of the target signal. To estimate the DoA, we present a maximum likelihood framework which takes the shadowing effect of the user's head on the received signals into account by modeling the relative transfer functions (RTFs) between the HAS's microphones. We propose three different RTF models which have different degrees of accuracy and individualization. Further, we show that the proposed DoA estimators can be formulated in terms of inverse discrete Fourier transforms (IDFTs) to evaluate the likelihood function computationally efficiently. We extensively assess the performance of the proposed DoA estimators for various DoAs, signal to noise ratios (SNRs), and in different noisy and reverberant situations. The results show that the proposed estimators improve the performance markedly over other recently proposed "informed" DoA estimator.**

***Index Terms*—Sound Source Localization, Direction of Arrival Estimation, Hearing Aid, Maximum Likelihood, Relative Transfer Function.**

I. Introduction

In realistic acoustic scenes, where several sound sources are present simultaneously, the auditory scene analysis (ASA) ability in humans allows them to focus deliberately on a sound source while suppressing the other irrelevant sound sources [1]. Sensorineural hearing loss degrades this ability [2], and hearing impaired listeners face difficulties in interacting with the environment. Hearing aid systems (HASs) may take some of these ASA responsibilities to restore the normal interactions of the hearing impaired users with the environment.

Sound source localization (SSL) is one of the main tasks in ASA, and different SSL approaches have been proposed for various applications, such as robotics [3], [4], video conferencing [5], surveillance [6], and hearing aids [7].

SSL strategies using microphone arrays can be generally categorized as[1]:

M. Farmani and Z.-H. Tan are with with Aalborg Univeristy, Department of Electronic Systems, Signal and Information Processing Section, 9220 Aalborg, Denmark (e-mail: mof@es.aau.dk; zt@es.aau.dk).

M. S. Pedersen is with Oticon A/S, 2765 Smørum, Denmark (e-mail: micp@oticon.com).

J. Jensen is with Aalborg Univeristy, Department of Electronic Systems, Signal and Information Processing Section, 9220 Aalborg, Denmark, and also with Oticon A/S, 2765 Smørum, Denmark (e-mail: jje@es.aau.dk; jesj@oticon.com).

Manuscript received MMM DD, YYYY; revised MMM DD, YYYY; accepted MMM DD, YYYY. Date of publication MMM DD, YYYY; date of current version MMM DD, YYYY.

[1]This is an extended version of the categorization proposed in [8, ch. 8].

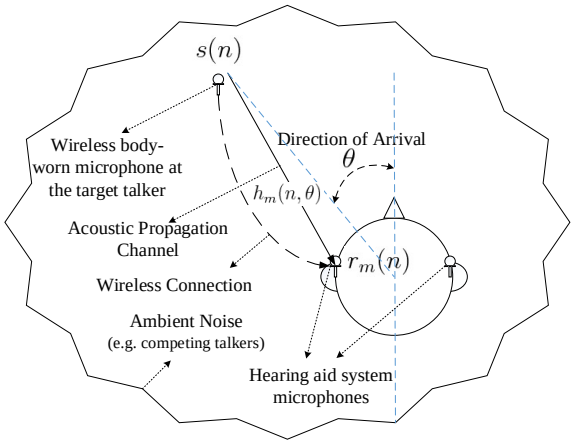


Fig. 1: An "informed" SSL scenario for a binaural hearing aid system using a wireless microphone. $r_m(n)$ is the noisy received sound at microphone m, $s(n)$ is the noise-free target sound emitted at the target location, and $h_m(n,\theta)$ is the acoustic channel impulse response between the target talker and microphone m. $s(n)$ is available at the HAS via the wireless connection, and the hearing aids are also connected to each other wirelessly. The goal is to estimate θ.

- Steered-beamformer-based (also called steered response power methods): the main idea of these methods is to steer a beamformer towards potential locations and look for a maximum in the output power [8, ch. 8],[9].
- High-resolution-spectral-estimation-based: these methods are based on the spatiospectral correlation matrix obtained from the microphones signals. Under certain assumptions, the sound source locations can be estimated from a lower-dimensional vector subspace embedded within the signal space spanned by the columns of the correlation matrix [10], [11].
- Time-difference-of-arrival (TDoA)-based: these methods first estimate a set of TDoAs of the signals reaching each pair of the microphones in the microphone array, then map the estimated TDoAs to an estimate of the sound source location using a mapping function [12], [13].
- Head-related-transfer-function (HRTF)-based: when the microphone array is mounted at the head and torso of humans or humanoid robots, the filtering effects of the head and torso on the incoming sounds can be used for SSL [4], [14]–[17].

Most existing SSL algorithms have been proposed for applications which are "uninformed" about the noise-free content

of the target sound, e.g. [3]–[7], [9]–[16]. However, recent HASs can employ a wireless microphone worn by the target talker to access an essentially noise-free version of the target signal emitted at the target talker's position [17]–[20]. Using a wireless microphone worn by the target talker introduces the "informed" SSL problem considered in this paper.

Fig. 1 depicts the situation considered in this paper. The HAS consists of two hearing aids (HAs) connected wirelessly and mounted on each ear of the user, and a wireless microphone worn by the target talker. The target signal $s(n)$ is emitted at the target location, propagates through the acoustic channel $h_m(n,\theta)$, and reaches microphone $m \in \{\text{left}, \text{right}\}$ of the binaural HAS. Due to additive environmental noise, the signal captured by microphone m, denoted by $r_m(n)$, is a noisy version of the target signal impinging on the microphone. The problem considered in this paper is to estimate the target signal Direction of Arrival (DoA) θ based on the wirelessly available target signal $s(n)$ and the noisy microphone signals $r_m(n)$. Estimating the target sound DoA in this system allows the HAS to enhance the spatial correctness of the acoustic scene presented to the HAS user, e.g. by imposing the corresponding binaural cues on the wirelessly received target sound [21].

The "informed" SSL problem for hearing aid applications was first investigated via a TDoA-based approach in [18]. The method proposed in [18] uses a cross-correlation technique to estimate the TDoA, then uses a sine law to map the estimated TDoA to a DoA estimate. The approach proposed in [18] has relatively low computational load, because it does not take the shadowing effect of the user's head and the ambient noise characteristics into account. Disregarding the head shadowing effect inevitably degrades the DoA estimation performance, especially when the target sound is located at the sides of the user's head, where the head shadowing has the highest impact on the received signals. Moreover, neglecting the ambient noise characteristics causes the estimator performance to be sensitive to the noise type.

In this paper, we present a maximum likelihood (ML) framework for "informed" SSL relying on the noise-free target signal and the ambient noise characteristics. Moreover, to improve the estimation accuracy, we consider the effects of the user's head on the received signals by modeling the direction-dependent relative transfer functions (RTFs) between the left and right microphones of the HAS. More precisely, we present three different RTF models: i) the free-field-far-field model, ii) the spherical-head model, and iii) the measured-RTF model. These models have different degrees of accuracy and individualization. Using the proposed ML framework and based on each of the RTF models, we propose an ML estimator for the target sound DoA. Moreover, besides the DoA, as a by-product, the proposed methods provide an ML estimate of the target signal propagation time between the target talker and the user. The propagation time can be easily converted to a distance estimate, which is an important information about the target location.

The free-field-far-field model and the spherical-head model have been proposed and used for informed DoA estimation in [19] and [20], respectively. In this paper, we introduce the measured-RTF model and its corresponding ML DoA estimator. Moreover, we provide a new unified presentation of all the models and investigate their performances extensively.

The idea of using measured RTFs for "uninformed" DoA estimation was already presented in [22]. The method proposed in [22] considers a narrow-band "uniformed" DoA estimation problem and solves it using a minimum mean square error approach. In contrast, our proposed estimator based on the measured-RTF model solves a wide-band "informed" DoA estimation problem using a ML approach. We show that formulating the "informed" DoA estimation problem as wide-band allows us to evaluate the proposed likelihood function in all frequency bins at once using inverse discrete Fourier transforms (IDFTs), which can be computed efficiently.

The general ML framework presented in this paper was first proposed in [17] for the informed SSL, using a database of measured HRTFs. The HRTF database was used to model the acoustic channel and the shadowing effect of a particular user's head. To estimate the DoA, the proposed method in [17], called MLSSL (maximum likelihood sound source localization), looks for the HRTF entry in the database which maximizes the likelihood of the observed microphone signals. MLSSL is markedly effective under severely noisy conditions when the detailed information of the user-specific HRTFs for different directions and different distances is available.

Compared with MLSSL, which is based on HRTFs, the proposed estimators in this paper are based on RTFs. In contrast to HRTFs, which are distance-dependent, RTFs are almost independent of the distance between the target talker and the user, especially in far-field situations [23]. The distance independency decreases the required memory and the computational overhead of the proposed estimators. This is because to estimate the DoA, the proposed estimators must search in a RTF database, which is only a function of DoA, while MLSSL searches in an HRTF database which is a function of both DoA and distance. Further, the proposed estimators in this paper can all be formulated in terms of IDFTs which can be computed efficiently.

The structure of this paper is as follows. In Sections II and III, the signal model and the ML framework are presented, respectively. Afterwards, in Section IV, different RTF models used for modeling the presence of the head are introduced. The proposed DoA estimators using the proposed RTF models and the ML framework are derived in Section V. In Section VI, the performance of the proposed estimators is evaluated and compared using experimental simulations. Lastly, we conclude the paper in Section VII.

II. Signal Model

Regarding Fig. 1, the noisy signal received at microphone $m \in \{\text{left}, \text{right}\}$ of the HAS is given by:

$$r_m(n) = s(n) * h_m(n,\theta) + v_m(n), \tag{1}$$

where $s(n)$, $h_m(n,\theta)$ and $v_m(n)$ are the noise-free target signal emitted at the target talker's position, the acoustic channel impulse response between the target talker and microphone m, and an additive noise component, respectively. Further, n is the discrete time index, and $*$ denotes the convolution operator.

Most state-of-the-art HASs operate in the short time Fourier transform (STFT) domain because it allows frequency dependent processing, computational efficiency and low latency algorithm implementations. Therefore, Let

$$R_m(l,k) = \sum_n r_m(n) w(n-lA) \mathrm{e}^{-\frac{j2\pi k}{N}(n-lA)},$$

denote the STFT of $r_m(n)$, where l and k are frame and frequency bin indexes, respectively, N is the discrete Fourier transform (DFT) order, A is the decimation factor, $w(n)$ is the windowing function, and $j=\sqrt{-1}$ is the imaginary unit. Similarly, let us denote the STFT of $s(n)$ and $v_m(n)$ by $S(l,k)$ and $V_m(l,k)$, respectively, which are defined analogously to $R_m(l,k)$. Moreover, let

$$\begin{aligned} H_m(k,\theta) &= \sum_n h_m(n,\theta)\mathrm{e}^{-\frac{j2\pi kn}{N}} \\ &= \alpha_m(k,\theta)\mathrm{e}^{-\frac{j2\pi k}{N}D_m(k,\theta)}, \end{aligned} \quad (2)$$

denote the discrete Fourier transform (DFT) of $h_m(n,\theta)$, where $\alpha_m(k,\theta)$ is a real positive number and denotes the frequency-dependent attenuation factor due to propagation effects, and $D_m(k,\theta)$ is the frequency-dependent propagation time measured in samples, from the target sound source to microphone m. Eq. (1) can be approximated in the STFT domain as:

$$R_m(l,k) = S(l,k)H_m(k,\theta) + V_m(l,k). \quad (3)$$

This approximation is known as the multiplicative transfer function (MTF) approximation [24], and its accuracy depends on the length and smoothness of the windowing function $w(n)$: the longer and the smoother the analysis window $w(n)$, the more accurate the approximation [24].

III. Maximum Likelihood Framework

To define the likelihood function, let us assume that the additive noise observed at the microphones follows a zero-mean circularly-symmetric complex Gaussian distribution:

$$\boldsymbol{V}(l,k) = \begin{bmatrix} V_{\text{left}}(l,k) \\ V_{\text{right}}(l,k) \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_v(l,k)), \quad (4)$$

where $\mathbf{C}_v(l,k)$ is the noise cross power spectral density (CPSD) matrix defined as $\mathbf{C}_v(l,k) = \mathrm{E}\{\boldsymbol{V}(l,k)\boldsymbol{V}^{\mathrm{H}}(l,k)\}$, where $\mathrm{E}\{.\}$ and superscript H represent the expectation and Hermitian transpose operators, respectively. Further, let us assume that the noisy observations are independent across frequencies (strictly speaking, this assumption holds when the correlation time of the signal is short compared with the frame length [25], [26]). Therefore, the likelihood function for frame l is defined by:

$$p(\underline{\underline{\mathbf{R}}}(l); \underline{\underline{\mathbf{H}}}(\theta)) = \prod_{k=0}^{N-1} \frac{1}{\pi^M \det\left[\mathbf{C}_v(l,k)\right]} \mathrm{e}^{\{-(\boldsymbol{Z}(l,k))^{\mathrm{H}}\mathbf{C}_v^{-1}(l,k)(\boldsymbol{Z}(l,k))\}}, \quad (5)$$

where $\det[.]$ denotes the matrix determinant, and

$$\begin{aligned} \underline{\underline{\mathbf{R}}}(l) &= [\ \boldsymbol{R}(l,0),\ \boldsymbol{R}(l,1),\ \cdots,\ \boldsymbol{R}(l,N-1)\ ], \\ \boldsymbol{R}(l,k) &= [\ R_{\text{left}}(l,k),\ R_{\text{right}}(l,k)\ ]^{\mathrm{T}},\ 0 \le k \le N-1, \\ \underline{\underline{\mathbf{H}}}(\theta) &= [\ \boldsymbol{H}(0,\theta),\ \boldsymbol{H}(1,\theta),\ \cdots,\ \boldsymbol{H}(N-1,\theta)\ ], \\ \boldsymbol{H}(k,\theta) &= [\ H_{\text{left}}(k,\theta),\ H_{\text{right}}(k,\theta)\ ]^{\mathrm{T}} \\ &= \begin{bmatrix} \alpha_{\text{left}}(k,\theta)e^{-j2\pi\frac{k}{N}D_{\text{left}}(k,\theta)} \\ \alpha_{\text{right}}(k,\theta)e^{-j2\pi\frac{k}{N}D_{\text{right}}(k,\theta)} \end{bmatrix}, \\ \boldsymbol{Z}(l,k) &= \boldsymbol{R}(l,k) - S(l,k)\boldsymbol{H}(k). \end{aligned}$$

To reduce the computational overhead, we consider the log-likelihood function and omit the terms independent of θ. Therefore, the reduced log-likelihood function is given by:

$$\mathcal{L}(\underline{\underline{\mathbf{R}}}(l); \underline{\underline{\mathbf{H}}}(\theta)) = \sum_{k=0}^{N-1} \{-(\boldsymbol{Z}(l,k))^{\mathrm{H}}\mathbf{C}_v^{-1}(l,k)(\boldsymbol{Z}(l,k))\}. \quad (6)$$

The ML estimate of θ is found by maximizing $\mathcal{L}$. However, to maximize $\mathcal{L}$ with respect to θ, we need to model and find the ML estimate of the parameters ($\alpha_{\text{left}}, D_{\text{left}}, \alpha_{\text{right}}$ and D_{right}) in $\underline{\underline{\mathbf{H}}}(\theta)$. Instead of estimating all the parameters separately, in the following, we present three different RTF models, which model and define the relations between the parameters in $\underline{\underline{\mathbf{H}}}(\theta)$ considering the influence of the user's head, and with different degrees of accuracy and individualization. These RTF models allow us to formulate $\mathcal{L}$ depending on the parameters of the transfer function between the target and only one, not both, of the microphones, while it also considers the head presence.

IV. Relative Transfer Function (RTF) models

The RTF between the left and the right microphones represents the filtering effect of the user's head. Moreover, this RTF defines the relation between the acoustic channels' parameters (the attenuations and the delays) corresponding to the left and the right microphones. An RTF is usually defined with respect to a reference microphone. Without loss of generality, let us consider the left microphone as the reference microphone; therefore, considering Eq. (2), the RTF at frequency bin k is defined by

$$\begin{aligned} \Psi(k,\theta) &= \frac{H_{\text{right}}(k,\theta)}{H_{\text{left}}(k,\theta)} \\ &= \Gamma(k,\theta)\mathrm{e}^{-j2\pi\frac{k}{N}\Delta D(k,\theta)}, \end{aligned}$$

where

$$\begin{aligned} \Gamma(k,\theta) &= \frac{\alpha_{\text{right}}(k,\theta)}{\alpha_{\text{left}}(k,\theta)}, \\ \Delta D(k,\theta) &= D_{\text{right}}(k,\theta) - D_{\text{left}}(k,\theta). \end{aligned}$$

We refer to $\Gamma(k,\theta)$ in dB as the inter-microphone level difference (IMLD), and to $\Delta D(k,\theta)$ in discrete time samples as the inter-microphone time difference (IMTD). In the following, three different models are presented for the RTF with different degrees of accuracy.

A. The free-field-far-field model

The free-field-far-field model $\mathbf{\Psi}_{\text{ff}}(\theta)$ is the simplest and the most straightforward model, which simply ignores the shadowing effect of the user's head and relies on a minimal number of user-related prior assumptions. In a free-field and far-field situation, the delay and the attenuation of an acoustic channel are frequency-independent. Therefore, using basic geometry rules, the IMTD can be formulated as [19]

$$\begin{aligned}\Delta D_{\text{ff}}(\theta) &= D_{\text{right}}(\theta) - D_{\text{left}}(\theta) \\ &= -\frac{a}{c}\sin(\theta), \end{aligned} \tag{7}$$

where a is the head diameter (or more precisely, the distance between the microphones) and c is the sound speed. It should be noted that $\theta = 0^\circ$ is exactly at the front of the user, and DoAs are defined clockwise with respect to 0°. Moreover, in a free-field and far-field situation, $\alpha_{\text{left}}(\theta) = \alpha_{\text{right}}(\theta)$, i.e.

$$\Gamma_{\text{ff}}(\theta) = \frac{\alpha_{\text{right}}(\theta)}{\alpha_{\text{left}}(\theta)} = 1. \tag{8}$$

Accordingly, the RTF in a free-field and far-field situation is given by:

$$\mathbf{\Psi}_{\text{ff}}(\theta) = [\Psi_{\text{ff}}(0,\theta), \Psi_{\text{ff}}(1,\theta), \cdots, \Psi_{\text{ff}}(N-1,\theta)]^{\text{T}}$$

where

$$\Psi_{\text{ff}}(k,\theta) = \mathrm{e}^{j2\pi\frac{k}{N}(\frac{a}{c}\sin(\theta))}, 0 \le k \le N-1.$$

B. The spherical-head model

For the spherical-head model $\mathbf{\Psi}_{\text{sp}}(\theta)$, we model the user's head as a rigid sphere. Even though the IMTD and the IMLD for a spherical head are generally frequency-dependent, here we assume that the IMTD and the IMLD, or more precisely the delays and the attenuations of the acoustic channels, are frequency-independent. The frequency-independency assumption keeps the model simple and decreases the computational load [20]. Moreover, our preliminary simulation results reveal that a frequency-dependent spherical-head model, which is a more accurate model with more parameters, does not necessarily provide more accurate DoA estimation. This is partly because the frequency-dependent model is over-fitted to the spherical head, while there is a mismatch between the spherical head and an actual head.

For a spherical head, the IMTD can be approximated by the Woodworth model [27, pp. 520-523]:

$$\Delta D_{\text{sp}}(\theta) = -\frac{a}{2c}\left(\theta + \sin(\theta)\right). \tag{9}$$

Moreover, to model the IMLD, we use the following expression inspired by the work in [28]:

$$20\log_{10}\Gamma_{\text{sp}}(\theta) = \gamma\sin(\theta), \tag{10}$$

where γ is a frequency-independent scaling factor. In [20], to find the best γ for the DoA estimation, we ran simulation using the theoretical HRTF of the spherical-head model proposed in [23]. The results showed that $\gamma = 6.5$ provides the best DoA estimation performance [20]. Therefore, the RTF for the spherical-head model is given by

$$\mathbf{\Psi}_{\text{sp}}(\theta) = [\Psi_{\text{sp}}(0,\theta), \Psi_{\text{sp}}(1,\theta), \cdots, \Psi_{\text{sp}}(N-1,\theta)]^{\text{T}},$$

where

$$\Psi_{\text{sp}}(k,\theta) = 10^{\frac{6.5\sin(\theta)}{20}}\mathrm{e}^{j2\pi\frac{k}{N}(\frac{a}{2c}(\theta+\sin(\theta)))}, 0 \le k \le N-1.$$

C. The measured-RTF model

The measured-RTF model $\mathbf{\Psi}_{\text{ms}}(\theta)$ is the most detailed and individualized model. This model uses a database of RTFs for different directions obtained from the corresponding HRTFs measured for the specific user. The measured RTF model is defined as

$$\mathbf{\Psi}_{\text{ms}}(\theta) = [\Psi_{\text{ms}}(0,\theta), \Psi_{\text{ms}}(1,\theta), \cdots, \Psi_{\text{ms}}(N-1,\theta)]^{\text{T}},$$

where

$$\Psi_{\text{ms}}(k,\theta) = \Gamma_{\text{ms}}(k,\theta)\mathrm{e}^{j\Phi_{\text{ms}}(k,\theta)}, 0 \le k \le N-1,$$

where

$$\Gamma_{\text{ms}}(k,\theta) = \frac{|\tilde{H}_{\text{right}}(k,\theta)|}{|\tilde{H}_{\text{left}}(k,\theta)|}, \tag{11}$$

$$\Phi_{\text{ms}}(k,\theta) = \angle\frac{\tilde{H}_{\text{right}}(k,\theta)}{\tilde{H}_{\text{left}}(k,\theta)}, \tag{12}$$

where $\tilde{H}_{\text{left}}(k,\theta)$ and $\tilde{H}_{\text{right}}(k,\theta)$ are the measured HRTFs[2] for the left and right microphones, respectively, and $|.|$ and $\angle$ denote the magnitude and the phase angle of a complex number, respectively.

V. Proposed DoA estimators

In this section, we derive DoA estimators based on each of the proposed RTF models (Section IV) using the ML framework (Section III). In the derivations, we denote the inverse of the noise CPSD matrix as

$$\mathbf{C}_v^{-1}(l,k) \equiv \begin{bmatrix} C_{11}(l,k) & C_{12}(l,k) \\ C_{21}(l,k) & C_{22}(l,k) \end{bmatrix}. \tag{13}$$

To derive the DoA estimators, we expand the reduced log-likelihood function $\mathcal{L}$ presented in Eq. (6). Let

$$\begin{aligned}\boldsymbol{\alpha}_{\text{left}}(\theta) &= [\alpha_{\text{left}}(0,\theta), \alpha_{\text{left}}(1,\theta), \cdots, \alpha_{\text{left}}(N-1,\theta)]^{\text{T}}, \\ \boldsymbol{D}_{\text{left}}(\theta) &= [D_{\text{left}}(0,\theta), D_{\text{left}}(1,\theta), \cdots, D_{\text{left}}(N-1,\theta)]^{\text{T}}, \\ \boldsymbol{\alpha}_{\text{right}}(\theta) &= [\alpha_{\text{right}}(0,\theta), \alpha_{\text{right}}(1,\theta), \cdots, \alpha_{\text{right}}(N-1,\theta)]^{\text{T}},\end{aligned}$$

and

$$\boldsymbol{D}_{\text{right}}(\theta) = [D_{\text{right}}(0,\theta), D_{\text{right}}(1,\theta), \cdots, D_{\text{right}}(N-1,\theta)]^{\text{T}}.$$

[2]Formally, an HRTF is defined as "a specific individuals left or right ear far-field frequency response, as measured from a specific point in the free field to a specific point in the ear canal" [29]. However, in this paper we relax this definition and use the term HRTF to describe the frequency response from a target source to the microphone of a hearing aid system.

The expansion of $\mathcal{L}$ is

$$\begin{aligned}\mathcal{L}\left(\underline{\underline{\mathbf{R}}}(l);\boldsymbol{\alpha}_{\text{left}}(\theta),\boldsymbol{D}_{\text{left}}(\theta),\boldsymbol{\alpha}_{\text{right}}(\theta),\boldsymbol{D}_{\text{right}}(\theta)\right)=\\ \sum_{k=1}^{N}2\alpha_{\text{left}}(k,\theta)C_{11}(l,k)R_{\text{left}}(l,k)S^{*}(l,k)\mathrm{e}^{\frac{j2\pi kD_{\text{left}}(k,\theta)}{N}}+\\ 2\alpha_{\text{left}}(k,\theta)C_{12}(l,k)R_{\text{right}}(l,k)S^{*}(l,k)\mathrm{e}^{\frac{j2\pi kD_{\text{left}}(k,\theta)}{N}}+\\ 2\alpha_{\text{right}}(k,\theta)C_{21}(l,k)R_{\text{left}}(l,k)S^{*}(l,k)\mathrm{e}^{\frac{j2\pi kD_{\text{right}}(k,\theta)}{N}}+\\ 2\alpha_{\text{right}}(k,\theta)C_{22}(l,k)R_{\text{right}}(l,k)S^{*}(l,k)\mathrm{e}^{\frac{j2\pi kD_{\text{right}}(k,\theta)}{N}}+\\ \left(\alpha^{2}_{\text{left}}(k,\theta)C_{11}(l,k)+\alpha^{2}_{\text{right}}(k,\theta)C_{22}(l,k)\right)|S(l,k)|^{2}+\\ 2\alpha_{\text{left}}(k,\theta)\alpha_{\text{right}}(k,\theta)C_{21}(l,k)|S(l,k)|^{2}\times\\ \mathrm{e}^{\frac{j2\pi k}{N}(D_{\text{right}}(k,\theta)-D_{\text{left}}(k,\theta))}.\end{aligned}\tag{14}$$

In the following, we aim to make $\mathcal{L}$ independent of all other parameters except θ, using the proposed RTF models.

A. The free-field-far-field model DoA estimator

As mentioned, in a free-field and far-field situation, the delays and the attenuations of acoustic channels are frequency independent. Based on Eqs. (7) and (8), $D_{\text{right}}(\theta)$ and $\alpha_{\text{right}}(\theta)$ can be written as functions of $D_{\text{left}}(\theta)$ and $\alpha_{\text{left}}(\theta)$, respectively:

$$\begin{aligned}D_{\text{right}}(\theta)&=\Delta D_{\text{ff}}(\theta)+D_{\text{left}}(\theta)\\&=-\frac{a}{c}\sin(\theta)+D_{\text{left}}(\theta),\\ \alpha_{\text{right}}(\theta)&=\Gamma_{\text{ff}}(\theta)\alpha_{\text{left}}(\theta)\\&=\alpha_{\text{left}}(\theta).\end{aligned}$$

Inserting these relations in Eq. (14), we arrive at the reduced log-likelihood function $\mathcal{L}(\underline{\underline{\mathbf{R}}}(l);\boldsymbol{\Psi}_{\text{ff}}(\theta),\alpha_{\text{left}}(\theta),D_{\text{left}}(\theta))$ which is independent of H_{right} parameters (i.e. $D_{\text{right}}(\theta)$ and $\alpha_{\text{right}}(\theta)$). To eliminate the dependency of $\mathcal{L}$ on $\alpha_{\text{left}}(\theta)$, we find the maximum likelihood estimate (MLE) of $\alpha_{\text{left}}(\theta)$ in terms of other parameters, and replace the result into $\mathcal{L}$. To do so, we solve $\frac{\partial\mathcal{L}}{\partial\alpha_{\text{left}}(\theta)}=0$, which leads to

$$\hat{\alpha}_{\text{left}}(\theta)=\frac{f_{\text{ff}}(\boldsymbol{\Psi}_{\text{ff}}(\theta),D_{\text{left}}(\theta))}{g_{\text{ff}}(\boldsymbol{\Psi}_{\text{ff}}(\theta))},\tag{15}$$

where

$$\begin{aligned}f_{\text{ff}}(\boldsymbol{\Psi}_{\text{ff}}(\theta),D_{\text{left}}(\theta))=\sum_{k=1}^{N}\Big(C_{11}(l,k)R_{\text{left}}(l,k)+\\ C_{12}(l,k)R_{\text{right}}(l,k)+\big(C_{21}(l,k)R_{\text{left}}(l,k)+\\ C_{22}(l,k)R_{\text{right}}(l,k)\big)\Psi^{*}_{\text{ff}}(k,\theta)\Big)\times\\ S^{*}(l,k)e^{\frac{j2\pi kD_{\text{left}}(\theta)}{N}},\end{aligned}\tag{16}$$

and

$$\begin{aligned}g_{\text{ff}}(\boldsymbol{\Psi}_{\text{ff}}(\theta))&=\sum_{k=1}^{N}\Big(C_{11}(l,k)+2C_{21}(l,k)\Psi^{*}_{\text{ff}}(k,\theta)+\\ &\qquad C_{22}(l,k)\Big)|S(l,k)|^{2}.\end{aligned}\tag{17}$$

Inserting $\hat{\alpha}_{\text{left}}$ into $\mathcal{L}$ gives us:

$$\mathcal{L}_{\text{ff}}(\underline{\underline{\mathbf{R}}}(l);\boldsymbol{\Psi}_{\text{ff}}(\theta),D_{\text{left}}(\theta))=\frac{f^{2}_{\text{ff}}\left(\boldsymbol{\Psi}_{\text{ff}}(\theta),D_{\text{left}}(\theta)\right)}{g_{\text{ff}}(\boldsymbol{\Psi}_{\text{ff}}(\theta))}.\tag{18}$$

From Eq. (16) it can be seen that for a given θ, $f_{\text{ff}}(\boldsymbol{\Psi}_{\text{ff}}(\theta),D_{\text{left}}(\theta))$ is an IDFT, which can be evaluated efficiently, with respect to $D_{\text{left}}(\theta)$, while $g_{\text{ff}}(\boldsymbol{\Psi}_{\text{ff}}(\theta))$ is a simple summation. Therefore, computing $\mathcal{L}_{\text{ff}}$ for a given θ results in a discrete-time sequence corresponding to different values of $D_{\text{left}}(\theta)$. Since θ is unknown, we consider a discrete set Θ of different θs, and compute $\mathcal{L}$ for each $\theta\in\Theta$ using an IDFT. Evaluating $\mathcal{L}$ for all $\theta\in\Theta$ results in a 2-dimensional discrete grid as a function of different values of θ and D_{left}. The MLEs of θ and D_{left} are then found from the global maximum:

$$\left[\hat{\theta}_{\text{ff}},\hat{D}_{\text{left}}\right]=\underset{\theta\in\Theta,D_{\text{left}}}{\arg\max}\,\mathcal{L}_{\text{ff}}(\underline{\underline{\mathbf{R}}}(l);\boldsymbol{\Psi}_{\text{ff}}(\theta),D_{\text{left}}(\theta)).\tag{19}$$

B. The spherical-head model DoA estimator

The derivation of the DoA estimator based on the spherical-head model is analogous to the free-field-far-field DoA estimator. We assume, as in the free-field-far-field model, that the delay and the attenuation of acoustic channels are frequency-independent, and we replace $D_{\text{right}}(\theta)$ and $\alpha_{\text{right}}(\theta)$ with functions of $D_{\text{left}}(\theta)$ and $\alpha_{\text{left}}(\theta)$, respectively, using Eqs. (9) and (10):

$$\begin{aligned}D_{\text{right}}(\theta)&=\Delta D_{\text{sp}}(\theta)+D_{\text{left}}(\theta)\\&=-\frac{a}{2c}\left(\sin(\theta)+\theta\right)+D_{\text{left}}(\theta),\end{aligned}\tag{20}$$

$$\begin{aligned}\alpha_{\text{right}}(\theta)&=\Gamma_{\text{sp}}(\theta)\alpha_{\text{left}}(\theta)\\&=10^{\frac{6.5\sin(\theta)}{20}}\alpha_{\text{left}}(\theta).\end{aligned}\tag{21}$$

Inserting Eqs. (20) and (21) into Eq. (14) makes $\mathcal{L}$ independent of $D_{\text{right}}(\theta)$ and $\alpha_{\text{right}}(\theta)$, i.e. we have $\mathcal{L}(\underline{\underline{\mathbf{R}}}(l);\boldsymbol{\Psi}_{\text{sp}}(\theta),\alpha_{\text{left}}(\theta),D_{\text{left}}(\theta))$. As for the free-field-far-field model, to find the MLE of $\alpha_{\text{left}}(\theta)$ as a function of the other parameters, we solve $\frac{\partial\mathcal{L}}{\partial\alpha_{\text{left}}(\theta)}=0$. The resulting MLE of $\alpha_{\text{left}}(\theta)$ can be expressed as

$$\hat{\alpha}_{\text{left}}(\theta)=\frac{f_{\text{sp}}(\boldsymbol{\Psi}_{\text{sp}}(\theta),D_{\text{left}}(\theta))}{g_{\text{sp}}(\boldsymbol{\Psi}_{\text{sp}}(\theta))},\tag{22}$$

where

$$\begin{aligned}f_{\text{sp}}(\boldsymbol{\Psi}_{\text{sp}}(\theta),D_{\text{left}}(\theta))=\sum_{k=1}^{N}\Big(C_{11}(l,k)R_{\text{left}}(l,k)+\\ C_{12}(l,k)R_{\text{right}}(l,k)+\big(C_{21}(l,k)R_{\text{left}}(l,k)+\\ C_{22}(l,k)R_{\text{right}}(l,k)\big)\Psi^{*}_{\text{sp}}(k,\theta)\Big)\times\\ S^{*}(l,k)e^{\frac{j2\pi kD_{\text{left}}(\theta)}{N}},\end{aligned}\tag{23}$$

and

$$\begin{aligned}g_{\text{sp}}(\boldsymbol{\Psi}_{\text{sp}}(\theta))&=\sum_{k=1}^{N}\Big(C_{11}(l,k)+2C_{21}(l,k)\Psi^{*}_{\text{sp}}(k,\theta)+\\ &\qquad\Gamma^{2}_{\text{sp}}(\theta)C_{22}(l,k)\Big)|S(l,k)|^{2}.\end{aligned}\tag{24}$$

Inserting Eq. (22) into $\mathcal{L}(\underline{\underline{\mathbf{R}}}(l);\boldsymbol{\Psi}_{\text{sp}}(\theta),\alpha_{\text{left}}(\theta),D_{\text{left}}(\theta))$ gives us:

$$\mathcal{L}_{\text{sp}}(\underline{\underline{\mathbf{R}}}(l);\boldsymbol{\Psi}_{\text{sp}}(\theta),D_{\text{left}}(\theta))=\frac{f^{2}_{\text{sp}}\left(\boldsymbol{\Psi}_{\text{sp}}(\theta),D_{\text{left}}(\theta)\right)}{g_{\text{sp}}(\boldsymbol{\Psi}_{\text{sp}}(\theta))}.\tag{25}$$

Again, it can be seen that $f_{\text{sp}}(\mathbf{\Psi}_{\text{sp}}(\theta), D_{\text{left}}(\theta))$ in Eq. (23) is an IDFT with respect to $D_{\text{left}}(\theta)$, and $g_{\text{sp}}(\mathbf{\Psi}_{\text{sp}}(\theta))$ is a simple summation for a given θ. As before, for a given θ, evaluating $\mathcal{L}_{\text{sp}}$ results in a discrete-time sequence corresponding to different discrete values of $D_{\text{left}}(\theta)$. Since θ is unknown, we consider a discrete set Θ of different θs, and compute $\mathcal{L}$ for each $\theta \in \Theta$ using an IDFT. The MLEs of θ and D_{left} are then found from the global maximum:

$$\left[\hat{\theta}_{\text{sp}}, \hat{D}_{\text{left}}\right] = \underset{\theta\in\Theta, D_{\text{left}}}{\arg\max}\, \mathcal{L}_{\text{sp}}(\underline{\underline{\mathbf{R}}}(l); \mathbf{\Psi}_{\text{sp}}(\theta), D_{\text{left}}(\theta)). \quad (26)$$

C. The measured-RTF model DoA estimator

In the measured-RTF model, we assume that a database Θ_{ms} of measured frequency-dependent RTFs, labeled by their corresponding directions, for the specific user, is available. The DoA estimator using this model is based on evaluating $\mathcal{L}$ for the different RTFs in Θ_{ms}. The DoA label of the RTF, which gives the highest likelihood is the MLE of the target DoA.

To evaluate $\mathcal{L}$ for each $\mathbf{\Psi}_{\text{ms}}(\theta) \in \Theta_{\text{ms}}$, we assume the parameters of the acoustic transfer function related to the "sunny" microphone is frequency independent. The "sunny" microphone is the microphone which is not in the "shadow" of the head, if we assume the sound is coming from the direction θ. To be more precise, when we evaluate $\mathcal{L}$ for $\mathbf{\Psi}_{\text{ms}}(\theta)$ corresponding to the directions on the left side of the head ($\theta \in [-90^\circ, 0^\circ]$), the acoustic transfer function parameters related to the left microphone, i.e. $\alpha_{\text{left}}(\theta)$ and $D_{\text{left}}(\theta)$, are assumed to be frequency independent. Similarly, when we evaluate $\mathcal{L}$ for $\mathbf{\Psi}_{\text{ms}}(\theta)$ corresponding to the directions on the right side of the head ($\theta \in (0^\circ, +90^\circ]$), the acoustic transfer function parameters related to the right microphone, i.e. $\alpha_{\text{right}}(\theta)$ and $D_{\text{right}}(\theta)$, are assumed to be frequency independent. Note that this evaluation strategy can be carried out in practice; it requires no prior knowledge about the true DoA.

This assumption about the "sunny" microphone is reasonable, because if the sound is really coming from direction θ, the signal received by the "sunny" microphone is almost unaltered by the head and torso of the user, i.e. this resembles a free-field situation. As shown below, this assumption allows us to use an IDFT for evaluation of $\mathcal{L}$. Note that this frequency-independency assumption is only related to the acoustic channel parameters from the target to one of the microphones. The RTFs between microphones are allowed to be frequency-dependent.

To evaluate $\mathcal{L}$ for $\mathbf{\Psi}_{\text{ms}}(\theta)$ where $\theta \in [-90^\circ, 0^\circ]$, let us replace $\alpha_{\text{right}}(k,\theta)$ and $D_{\text{right}}(k,\theta)$ in $\mathcal{L}$ with functions of $D_{\text{left}}(\theta)$ and $\alpha_{\text{left}}(\theta)$, respectively:

$$\begin{aligned} \alpha_{\text{right}}(k,\theta) &= \Gamma_{\text{ms}}(k,\theta)\alpha_{\text{left}}(\theta), & (27)\\ D_{\text{right}}(k,\theta) &= \Delta D_{\text{ms}}(k,\theta) + D_{\text{left}}(\theta) \\ &= \frac{-N}{2\pi k}\left(\Phi_{\text{ms}}(k,\theta) + 2\pi\rho\right) + D_{\text{left}}(\theta), & (28)\end{aligned}$$

where ρ is a phase unwrapping factor. This makes $\mathcal{L}$ independent of H_{right} parameters. Afterwards, as before, to make $\mathcal{L}$ independent of $\alpha_{\text{left}}(\theta)$, we find the MLE of $\alpha_{\text{left}}(\theta)$ as functions of other parameters in $\mathcal{L}$ by solving $\frac{\partial \mathcal{L}}{\partial \alpha_{\text{left}}(\theta)} = 0$. The obtained MLE of $\alpha_{\text{left}}(\theta)$ is:

$$\hat{\alpha}_{\text{left}}(\theta) = \frac{f_{\text{ms,left}}(\mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{left}}(\theta))}{g_{\text{ms,left}}(\mathbf{\Psi}_{\text{ms}}(\theta))}, \quad (29)$$

where

$$\begin{aligned} f_{\text{ms,left}}(\mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{left}}(\theta)) = \sum_{k=1}^{N} \Big(& C_{11}(l,k)R_{\text{left}}(l,k) + \\ & C_{12}(l,k)R_{\text{right}}(l,k) + \big(C_{21}(l,k)R_{\text{left}}(l,k) + \\ & C_{22}(l,k)R_{\text{right}}(l,k)\big)\Psi^*_{\text{ms}}(k,\theta)\Big) \times \\ & S^*(l,k)e^{\frac{j2\pi k D_{\text{left}}(\theta)}{N}}, \end{aligned} \quad (30)$$

and

$$\begin{aligned} g_{\text{ms,left}}(\mathbf{\Psi}_{\text{ms}}(\theta)) = \sum_{k=1}^{N} \Big(& C_{11}(l,k) + 2C_{21}(l,k)\Psi^*_{\text{ms}}(k,\theta) + \\ & \Gamma^2_{\text{ms}}(\theta)C_{22}(l,k)\Big)|S(l,k)|^2. \end{aligned} \quad (31)$$

Substituting $\hat{\alpha}_{\text{left}}(\theta)$ in $\mathcal{L}$ leads to

$$\mathcal{L}_{\text{ms,left}}(\underline{\underline{\mathbf{R}}}(l); \mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{left}}(\theta)) = \frac{f^2_{\text{ms,left}}(\mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{left}}(\theta))}{g_{\text{ms,left}}(\mathbf{\Psi}_{\text{ms}}(\theta))}.$$

Analogously, to evaluate $\mathcal{L}$ for $\mathbf{\Psi}_{\text{ms}}(\theta)$ where $\theta \in (0^\circ, +90^\circ]$, if we replace $\alpha_{\text{left}}(k,\theta)$ and $D_{\text{left}}(k,\theta)$ in $\mathcal{L}$ with functions of $\alpha_{\text{right}}(\theta)$ and $D_{\text{right}}(\theta)$, respectively, and go through the similar process, we end up with

$$\mathcal{L}_{\text{ms,right}}(\underline{\underline{\mathbf{R}}}(l); \mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{right}}(\theta)) = \frac{f^2_{\text{ms,right}}(\mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{right}}(\theta))}{g_{\text{ms,right}}(\mathbf{\Psi}_{\text{ms}}(\theta))},$$

where

$$\begin{aligned} f_{\text{ms,right}}(\mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{right}}(\theta)) = \sum_{k=1}^{N} \Big(& C_{21}(l,k)R_{\text{left}}(l,k) + \\ & C_{22}(l,k)R_{\text{right}}(l,k) + \big(C_{11}(l,k)R_{\text{left}}(l,k) + \\ & C_{12}(l,k)R_{\text{right}}(l,k)\big)(\Psi^*_{\text{ms}})^{-1}(k,\theta)\Big) \times \\ & S^*(l,k)e^{\frac{j2\pi k D_{\text{right}}(\theta)}{N}}, \end{aligned} \quad (32)$$

and

$$\begin{aligned} g_{\text{ms,right}}(\mathbf{\Psi}_{\text{ms}}(\theta)) = \sum_{k=1}^{N} \Big(& C_{22}(l,k) + 2C_{12}(l,k)(\Psi^*_{\text{ms}}(k,\theta))^{-1} + \\ & \Gamma^{-2}_{\text{ms}}(\theta)C_{11}(l,k)\Big)|S(l,k)|^2. \end{aligned} \quad (33)$$

Regarding Eqs. (30) and (32), $f_{\text{ms,left}}(\mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{left}}(\theta))$ and $f_{\text{ms,right}}(\mathbf{\Psi}_{\text{ms}}(\theta), D_{\text{right}}(\theta))$ can be seen to be IDFTs with respect to $D_{\text{left}}(\theta)$ and $D_{\text{right}}(\theta)$, respectively. Therefore, for a given θ, evaluating $\mathcal{L}_{\text{ms,left}}$ or $\mathcal{L}_{\text{ms,right}}$ results in a discrete-time sequence corresponding to different discrete values of $D_{\text{left}}(\theta)$ or $D_{\text{right}}(\theta)$. Therefore, evaluating $\mathcal{L}$ for all $\mathbf{\Psi}_{\text{ms}}(\theta) \in \Theta_{\text{ms}}$ results in a 2-dimensional discrete grid. The MLEs of θ and D_{left} or D_{right} are then found from the global maximum:

$$\left[\hat{\theta}_{\text{ms}}, \hat{D}\right] = \underset{\mathbf{\Psi}_{\text{ms}}(\theta)\in\Theta_{\text{ms}}, D}{\arg\max}\, \mathcal{L}_{\text{ms}}(\underline{\underline{\mathbf{R}}}(l); \mathbf{\Psi}_{\text{ms}}(\theta), D(\theta)), \quad (34)$$

where

$$\mathcal{L}_{\text{ms}}(\underline{\underline{\mathbf{R}}}(l);\boldsymbol{\Psi}_{\text{ms}}(\theta),D(\theta)) = \begin{cases}\mathcal{L}_{\text{ms,left}}(\underline{\underline{\mathbf{R}}}(l);\boldsymbol{\Psi}_{\text{ms}}(\theta),D_{\text{left}}(\theta)) & ,\theta\in[-90^\circ,0^\circ]\\ \mathcal{L}_{\text{ms,right}}(\underline{\underline{\mathbf{R}}}(l);\boldsymbol{\Psi}_{\text{ms}}(\theta),D_{\text{right}}(\theta)) & ,\theta\in(0^\circ,+90^\circ]\end{cases}.$$

VI. Simulation Results

In this section, we evaluate the performance of the estimators in simulation experiments. Specifically, we study the effects of the target sound DoA θ, the signal-to-noise ratio (SNR), the frame length, the noise type and the reverberation.

A. Implementation

The simulation parameters are generally as follows: the sampling frequency is 16 kHz, the DFT order $N = 512$, $w(n)$ is a Hamming window, the length of the window $w(n)$ is the same as the DFT order N, $A = \frac{N}{2}$, and the microphone distance $a = 16.4\,\text{cm}$. Moreover, to evaluate the likelihood functions, the noise CPSD matrix $\mathbf{C}_v(l,k)$ must be known. In the following, the procedure for estimating $\mathbf{C}_v(l,k)$ is outlined.

1) Estimating the noise CPSD matrix: to estimate $\mathbf{C}_v(l,k)$ in practice, we use $S(l,k)$, which is available at the HAS, as a voice activity detector. Specifically, access to $S(l,k)$ allows us to determine the time-frequency regions in $\boldsymbol{R}(l,k)$, where the target speech is essentially absent, and to adaptively estimate $\mathbf{C}_v(l,k)$ via recursive averaging [17], [30].

Alg. 1 shows the procedure for estimating $\mathbf{C}_v(l,k)$. If the difference between the maximum energy $S_{max}(k)$ in frequency bin k of the target signal observed so far and the energy of $S(l,k)$ in dB is larger than a certain threshold δ_{th}, we assume the target signal to be absent in frame l and frequency bin k. Hence, $\boldsymbol{R}(l,k)$ is noise dominated in this time-frequency region. Therefore, the estimate of $\mathbf{C}_v(l,k)$ is updated via exponential smoothing with a smoothing factor $0<\eta<1$. On the other hand, if the difference is smaller than the threshold δ_{th}, the target signal is assumed to be present in $\boldsymbol{R}(l,k)$. Therefore, the estimate of $\mathbf{C}_v$ is not updated, i.e. $\mathbf{C}_v(l,k)=\mathbf{C}_v(l-1,k)$. Finally, we update $S_{max}(k)$ if needed, or use a forgetting factor $0<\beta<1$ to adapt $S_{max}(k)$ with the possible changes in the target signal over time, e.g. if the target talker has changed, or if the target talker stops speaking. We use $\delta_{\text{th}} = 25\,\text{dB}$, $\eta = 0.9$ and $\beta = 0.95$ in the implementation.

Algorithm 1: Estimation of $\mathbf{C}_v(l,k)$

Input : $\boldsymbol{R}(l,k)$, $S(l,k)$
Output: $\mathbf{C}_v(l,k)$

```
if S_max(k) − 20 log10 |S(l,k)| > δ_th then
    /* Target signal is almost absent */
    C_v(l,k) = η R(l,k) * R(l,k)^H + (1−η) C_v(l−1,k);
else
    C_v(l,k) = C_v(l−1,k);
end
if S_max(k) < 20 log10 |S(l,k)| then
    S_max(k) = 20 log10 |S(l,k)|
else
    S_max(k) = S_max(k) + 10 log10(β)
end
```

B. Acoustic setup

To simulate real world scenarios, we use the database of head related impulse responses (HRIRs) and binaural room impulse responses, provided by [31]. We use a subset of the database for the frontal-horizontal plane $\theta\in\Theta=\{-85^\circ,-80^\circ,\cdots,+85^\circ\}$ measured with behind-the-ear (BTE) hearing aids mounted behind the ears of a head-and-torso simulator (HATS). We consider only the frontal-horizontal plane because in practice, the target talker is usually located at the front of the user. Moreover, because of the head symmetry and the microphone locations, the estimators suffer from front-back confusions, as humans do [32]. Therefore, considering only the frontal plane allows to avoid the influence of the front-back confusions on the estimators performance. To simulate a signal from a particular position, we convolve the signal with the corresponding impulse response.

As a target signal, we consider a four-minute speech signal composed of two male and two female voices from the TSP database [33]. To evaluate the performance of the estimators in different noisy situations, we consider four different noise types: car-interior noise, speech-shaped noise, large-crowd noise, and bottling-factory-hall noise. These noise types cover noise signals with low-frequency content (the car-interior noise), high-frequency content (the bottling-factory-hall noise), stationary noises (the speech-shaped noise) and non-stationary noises (the large-crowd noise). The long-term power spectrum of the target signal emitted at the target position and the noise signals received at the left microphone are depicted in Fig. 2. To simulate a large-crowd noise field, we play back simultaneously 72 different speech signals from 72 different positions, which are uniformly distributed on a circle in the horizontal plane centered at the HATS. Similarly, for the speech-shaped noise and the bottling-factory-hall noise, we play back different realizations of the considered noise signal from all 72 considered positions simultaneously. The car-interior noise field, however, is a binaural recording measured by BTE hearing aids mounted behind the ears of a HATS placed on the passenger seat of a car driving in a city. The wide-band SNR, to be reported for each simulation experiment, is expressed relative to the left-ear microphone signals.

C. Performance metric

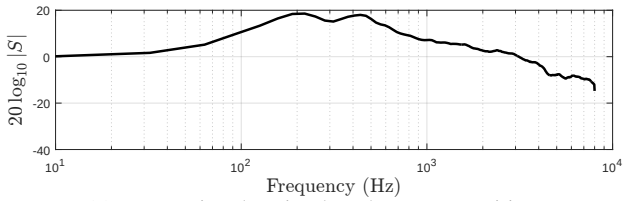
As a performance metric, we use the mean absolute error (MAE) of the DoA estimation, given by:

$$\text{MAE} = \frac{1}{\text{L}}\sum_{j=1}^{\text{L}}|\theta-\hat{\theta}_j|, \tag{35}$$

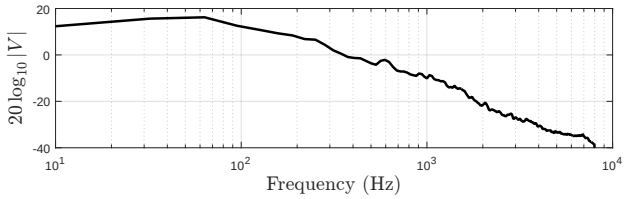
where $\hat{\theta}_j$ is the estimated DoA for the j^{th} frame of the signal, and L is the number of target-active frames (the target-inactive frames are disregarded).

D. Competing methods

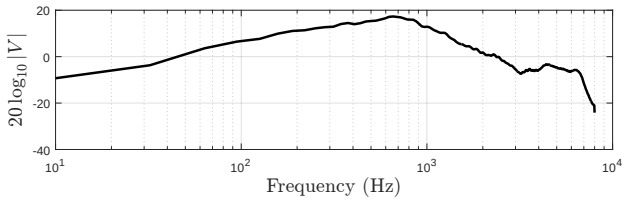
We compare the proposed estimators with the methods proposed in [18] and [17]. As outlined in Section I, the method



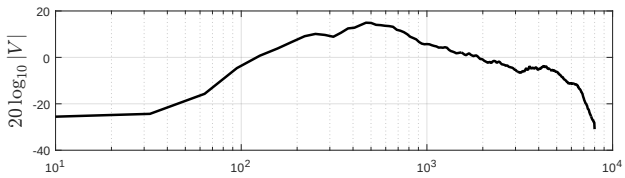
(a) Target signal emitted at the target position.



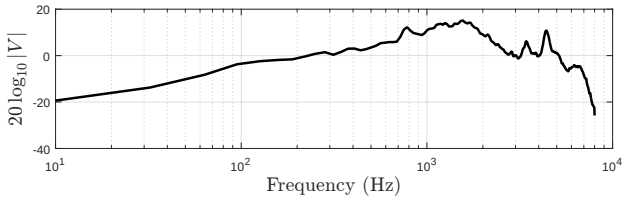
(b) Car-interior noise at the left microphone.



(c) Speech-shaped noise at the left microphone.

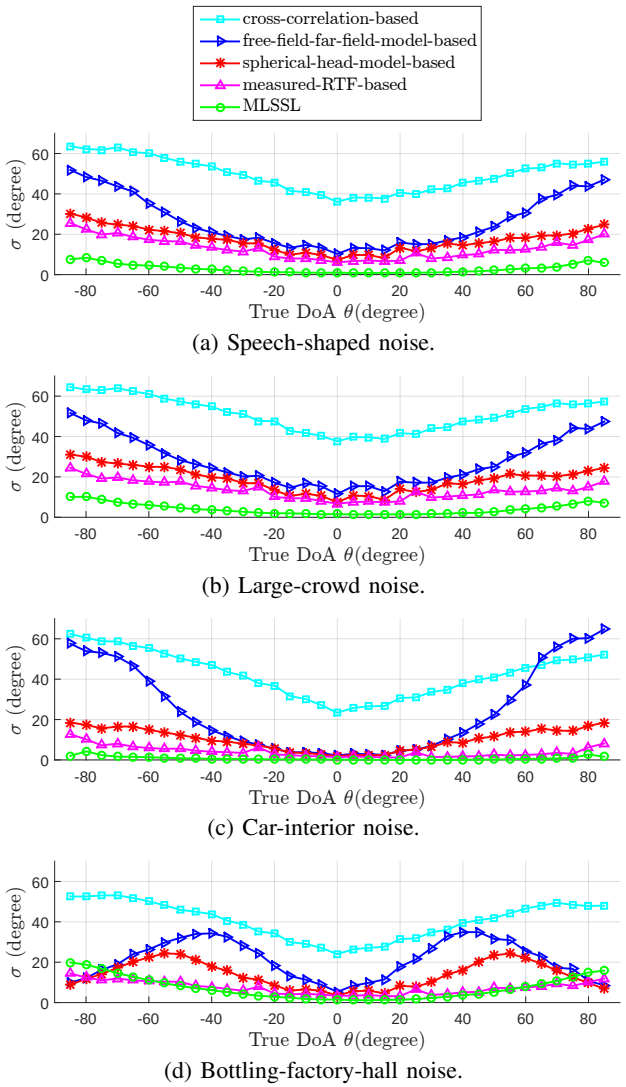


(d) Large-crowd noise at the left microphone.



(e) Bottling-factory-hall noise at the left microphone.

Fig. 2: Long-term power spectrum of the signals.



(a) Speech-shaped noise.

(b) Large-crowd noise.

(c) Car-interior noise.

(d) Bottling-factory-hall noise.

Fig. 3: Performance as a function of θ in an anechoic situation at SNR of 0 dB for different noise fields. The distance between the user and the target source is 300 cm. The HRTF database used for generation of the target signal is identical to the HRTF database used by MLSSL and the HRTFs used to build the measured-RTF model.

proposed in [18], which we refer to as the cross-correlation-based method, is simple because it does not take the ambient noise characteristics and the head shadowing effect into account. However, to model the curved path between the microphones, the distance between the microphones is assumed to be 25.2 cm, which is larger than the actual microphones distance. This particular distance is used because it leads to the best performance [18]. On the other hand, the method proposed in [17], called MLSSL, is a complex method. It takes the ambient noise characteristics into account by a maximum likelihood approach, and it exploits the details of the head shadowing effect via a database of HRTFs. In the MLSSL implementation, we use the same measured HRTF database, which is used to build the measured-RTF model.

E. Results and discussions

1) Influence of the target DoA: Fig. 3 compares the performance of the DoA estimators as a function of θ in an anechoic situation at SNR of 0 dB in different noise fields. As can be seen, the performance of all the estimators proposed in this paper are markedly more accurate than the performance of the cross-correlation-based method proposed in [18].

The poor performance of the cross-correlation-based method can be partly explained by the fact that the conventional cross-correlation technique is a maximum-likelihood optimal TDoA estimator for the situation, where the noise is white and Gaussian [34]. However, the frequency characteristics of the considered noise fields, shown in Fig. 2, are different from a white noise. This difference degrades considerably the performance of the cross-correlation-based method.

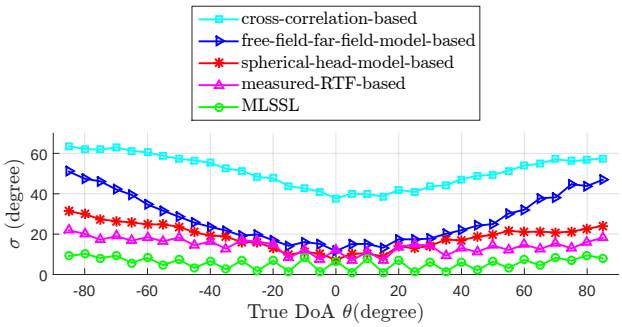


Fig. 4: Performance as a function of θ in an anechoic situation at SNR 0 dB in the large-crowd noise field. The HRTF database used by MLSSL and the measured-RTF database do not have any entries for every other considered θs for simulation.

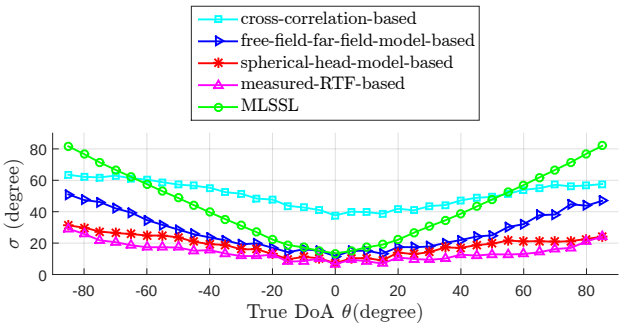


Fig. 5: Performance as a function of θ in an anechoic situation at SNR 0 dB in the large-crowd noise field. The distance between the user and the target source is 300 cm. The HRTF database used by MLSSL and the HRTF database used to build the measured-RTF model are for the case where the target is 80 cm away from the user.

Among the estimators proposed in this paper, the estimator based on the free-field-far-field model has the worst performance because it does not consider the shadowing effect of the user's head. In contrast, the spherical-head-model-based estimator models the head shadowing effect and improves the performance of the DoA estimation significantly, especially when the target is located at the sides of the HATS ($\theta \approx \pm 85°$), because this is where the shadowing effect of the head has the highest impact. When the user-specific, measured RTFs are available, even better performance can be achieved, because the influence of the head and torso is modeled more accurately.

Finally, as can be seen in Fig. 3, the performance of MLSSL is better than the performance of the measured-RTF-based estimator. This is because the exact HRTFs corresponding to the target locations are in the database searched by MLSSL, i.e. a highly idealized situation. Frequency-dependent HRTFs, as used in MLSSL, represent the acoustic transfer functions more accurately than the signal model used in the measured-RTF-based method, where the parameters of the acoustic channel between the target source and the microphone which is not in the head "shadow" are assumed to be frequency independent.

Another point to be made from Fig. 3 is that, similar to the sound source localization performance of humans [32], the general performance of the estimators when the target is at the sides (i.e. $\theta \approx \pm 90$) is worse than when the target is at the front ($\theta \approx 0°$). This is because the HRTFs (RTFs) corresponding to the front vary stronger within a certain angular range than the HRTFs (RTFs) corresponding to the sides [35]. In other words, when $\theta \in [-90°, -75°]$ or $\theta \in [75°, 90°]$, it is more probable to confuse the true HRTF (RTF) with the nearby HRTFs (RTFs).

2) Influence of the resolution of the databases: In practice, none of the entries in the HRTF database used by MLSSL or none of the entries in the RTF database used by the measured-RTF-based method can be expected to represent the actual DoA or distance of the target. Here, we investigate the performance of the estimators in these situations.

First, let us consider situations where the exact θ are not represented in the databases. To assess the performance of MLSSL and the measured-RTF-based estimator in these situations, we constructed reduced databases by eliminating every other entry from the MLSSL HRTF database and from the measured-RTF-model database. In other words, there is no entry in the databases for half of the considered target θs. Fig. 4 shows the performance of the estimators in this case. Comparing Fig. 4 with Fig. 3b shows that when the exact θ is not in the databases, the performance of MLSSL and the measured-RTF-based estimator degrade, as expected. However, most often, they succeed in finding the database entry closest to the target θ.

Next, we consider situations where the HRTFs corresponding to the actual distance between the target and the user are not in the database searched by MLSSL or in the HRTF database used to build the measured-RTF model. Fig. 5 shows the performance in such a situation, where the actual distance between the user and the target is 300 cm, but the employed HRTF database is for the case where the target is 80 cm away from the user (the database contains HRTFs for all the considered directions). It can be seen that the performance of MLSSL degrades dramatically in this situation: MLSSL is extremely sensitive to these HRTF mismatches. However, when the same HRTF database is used to build the measured-RTF model, the performance of the measured-RTF-based method degrades only slightly compared with Fig. 3. This robustness to the distance mismatches is because the measured RTFs are relatively distance independent. Therefore, the database used by the measured-RTF-based method can be just a function of the DoA, leading to a significant reduction of both memory and search complexity over the MLSSL method.

3) Influence of SNR: The SNR is another factor which generally influences the estimation performance. Fig. 6 shows the performance for different SNRs in terms of the MAE averaged over all considered θs in an anechoic situation in a large-crowd noise field. As expected, the higher the SNR, the better the performance. Moreover, as can be seen, the general performance order of Fig. 3 remains at different SNRs; however, the performance of the proposed measured-RTF-based method is almost the same as the performance of the MLSSL at high SNRs.

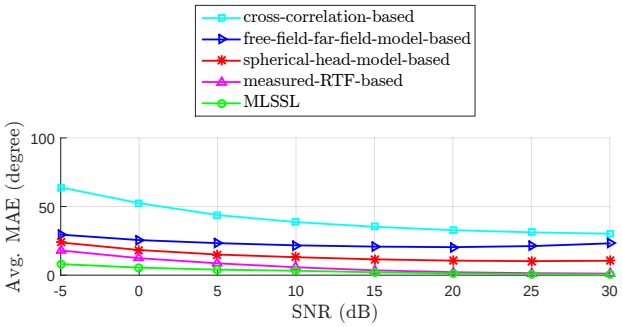


Fig. 6: Performance as a function of SNR in the same situation as in Fig. 3. The MAE is averaged over all considered θs.

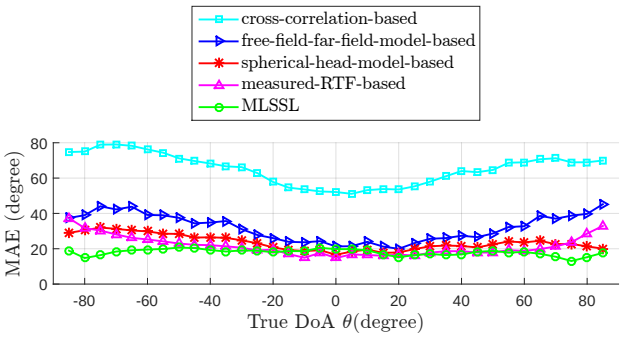


Fig. 7: Performance as a function of θ in a reverberant office with a reverberation time T_{60} of around 500 ms at SNR of 0 dB. The target is one meter away from the user. The HRTF database used by MLSSL, and the HRTFs used to build the measured-RTF model are "dry" and "clean" HRTFs for the case where the target is 80 cm away.

4) Influence of reverberation: Many speech communication situations occur indoor, where reverberation exists. Therefore, it is important to study the impact of reverberation on the performance of the estimators. Fig. 7 shows the performance of the DoA estimators as a function of θ in a reverberant office ($T_{60} \approx 500$ ms) at SNR of 0 dB in a large-crowd noise field. In contrast to Fig. 3, performance of all the estimators is reduced because none of them directly considers and models the reverberation. Even though, on average, the general performance order of Fig. 3 remains, the performance of the spherical-head-model-based method, the measured-RTF method and the MLSSL method approach each other. This is partly because the available "clean" HRTF database used by MLSSL and used to build the measured-RTF model are for the case where the target is 80 cm away while the actual distance of the target is 100 cm in the simulations.

5) Influence of the window length: Another factor which influences the performance of the estimators is the window (frame) length. Generally, at the cost of higher computational overhead and longer algorithmic delay, longer window lengths must lead to better performance because: 1) greater window lengths provide more observations, which reduces the variance of the estimates in a noisy situation, 2) the MTF approximation (Eq. 3) depends on the window length: the greater the window length, the better the approximation [24], and 3) greater window lengths strengthen the assumption that DFT coefficients are independent across frequencies (this assumption was used to write the simplified likelihood function in Eq. (5)). On the other hand, increasing the window length may violate the assumption implicitly made in Eq. (5) that signals are stationarity within a window duration.

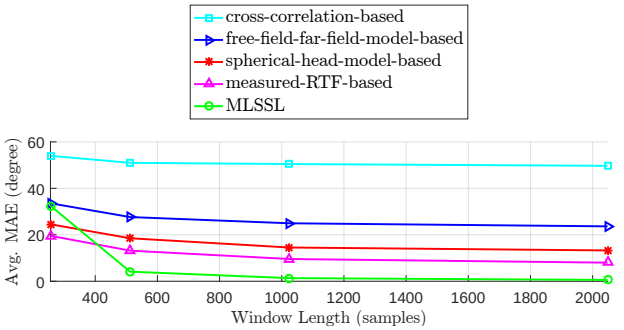


Fig. 8: Performance as a function of N in the same condition as in Fig. 3. The MAE is averaged over all considered θs.

Fig. 8 shows the performance of the DoA estimators as a function of window length. The results are consistent with the expectations: greater window lengths lead to better performance. Interestingly, even though MLSSL has better performance at longer window lengths, its performance is apparently very sensitive to smaller window lengths and deteriorates dramatically compared with the proposed estimators performance.

6) Influence of non-individualized HRTF databases: MLSSL and the measured-RTF-based method rely on HRTF databases measured for a specific user, and so far, we have presented their performance when user-specific databases are available. In some situations, measuring HRTFs for each user is impractical; however, it is possible to measure the HRTFs for a HATS beforehand. Therefore, in this part, we would like to compare the performance of the estimators in two different cases: 1) individualized: user-specific HRTF databases are available. 2) non-individualized: user-specific HRTF databases are not available; however, the corresponding databases measured for a HATS is available.

For the simulation, we use the HRTFs measured for binaural BTE hearing aids for five different persons (three males and two females) and a HATS. The HRTFs are measured in an anechoic situation for the frontal-horizontal plane.

Fig. 9 shows the performance of the estimators for the considered cases at an SNR of 0 dB in the large-crowd noise field. As can be seen, MLSSL is very sensitive to the mismatches in user-specific HRTF database. It has the best performance for all the users (subjects) when the user-specific HRTFs are available (the individualized case), but its performance degrades significantly when the HATS database is used for the DoA estimation (the non-individualized case). On the other hand, the measured-RTF-based method is much less sensitive. Overall, the measured-RTF-based method performs markedly better than MLSSL in the non-individualized case (when only the HATS database is available for the DoA estimation). The performance of the measured-RTF-based method in the non-individualized case

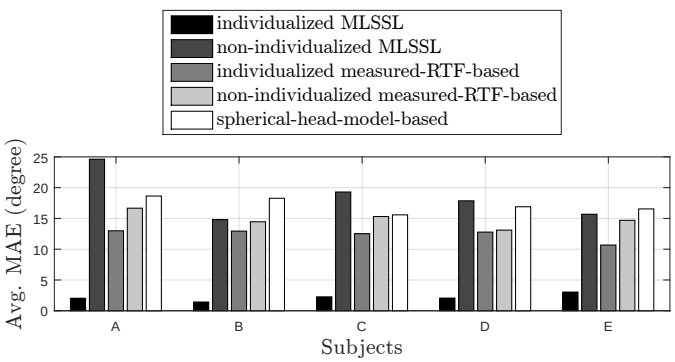


Fig. 9: Influence of non-individualized HRTF databases on the DoA estimators. The SNR is 0 dB in the large-crowd noise field. The MAE is averaged over all considered θs.

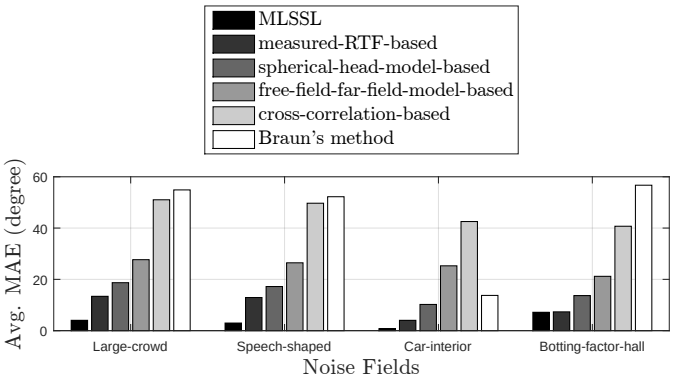


Fig. 10: Comparison of the "informed" DoA estimators with an "uninformed" DoA estimator proposed in [22], in different noise fields. The simulation was done in the same conditions as in Fig. 3. The MAE is averaged over all considered θs.

is also better than the spherical-head-model-based method, which does not depend on any user-specific databases.

7) Informed estimator vs. uninformed estimator: To demonstrate the benefits of access to the noise-free target signal, here we compare the performance of the proposed "informed" DoA estimators with the performance of a recently developed "uninformed" DoA estimator [22], which we refer to as Braun's method. As mentioned in Section I, Braun's method is a narrow-band estimator based on the measured-RTF model for the "uninformed" DoA estimation problem, i.e. where the clean target signal is not available. Regarding Eq. (3), it has been shown in [22] that the minimum mean square error (MMSE) estimator of the RTF between the two microphones at a particular frequency bin is given by:

$$\hat{\Psi}_{i,j}(k,\theta) = \frac{\phi_{R_{i,j}} - \phi_{V_{i,j}}}{\phi_{R_{j,j}} - \phi_{V_{j,j}}}, \tag{36}$$

where i and j are microphone indexes, $\phi_{R_{i,j}} = \mathrm{E}\{R_i(l,k)R_j^*(l,k)\}$ and $\phi_{V_{i,j}} = \mathrm{E}\{V_i(l,k)V_j^*(l,k)\}$. To make the estimate more robust, Braun's method averages the RTF estimate over the microphone index permutations, i.e.

$$\bar{\Psi}_{i,j}(k,\theta) = \frac{1}{2}\left\{\hat{\Psi}_{i,j}(k,\theta) + \hat{\Psi}_{j,i}^{-1}(k,\theta)\right\}. \tag{37}$$

Regarding the measured-RTF model Θ_{ms}, Braun's method estimates the DoA θ of the target signal at a particular frequency bin by

$$\hat{\theta}_{\text{Braun}} = \underset{\Psi_{\text{ms}}(k,\theta)\in\Theta_{\text{ms}}}{\arg\min} \sum_{i,j\in\mathcal{M}} W_{i,j}|\bar{\Psi}_{i,j}(k,\theta) - \Psi_{\text{ms}}(k,\theta)|, \tag{38}$$

where the set $\mathcal{M}$ contains all microphone pair combinations, and $W_{i,j}$ is a weighting factor for the $\{i,j\}$-th pair. In our setup, because we only have one microphone pair, we drop $W_{i,j}$ and consider $i =$ right and $j =$ left. Moreover, because the target in our problem is at the same position in all frequency bins, we modify the cost function as follows, to integrate the information of all frequency bins:

$$\hat{\theta}_{\text{Braun}} = \underset{\boldsymbol{\Psi}_{\text{ms}}(\theta)\in\Theta_{\text{ms}}}{\arg\min} \sum_{k=0}^{N-1} |\bar{\Psi}_{i,j}(k,\theta) - \Psi_{\text{ms}}(k,\theta)|. \tag{39}$$

To implement Braun's method, we used the same measured-RTF model as used by the proposed "informed" measured-RTF-based estimator. Moreover, as proposed in [22], to estimate $\phi_{R_{i,j}}$, a recursive averaging technique with a time constant of 50 ms was used. Finally, to estimate $\phi_{V_{i,j}}$ used in Braun's method, we use the estimation of $\mathbf{C}_v$ outlined in Section VI-A.

Fig. 10 shows the performance of the proposed "informed" DoA estimators vs. Braun's method. Clearly, the proposed DoA estimators, which have access to the noise-free target signal, perform markedly better than Braun's method, which does not have access to the noise-free signal. Moreover, in large-crowd noise, speech-shaped noise and bottling-factory-hall noise fields, the cross-correlation-based estimator, which is an "informed" estimator with low computational complexity, performs slightly better than Braun's method, which has relatively higher computational overhead. However, the estimation error of Braun's method significantly decreases in the car-interior noise, which is relatively stationary low frequency noise (c.f. Fig. 2b). At the cost of higher computational complexity, the performance of Braun's method could be improved to some extent by measuring the positive definiteness of $\mathbf{Q}(l,k) = \mathrm{E}\left\{\boldsymbol{R}(l,k)\boldsymbol{R}^{\mathrm{H}}(l,k)\right\} - \mathbf{C}_v(l,k)$, before subtracting the correlations in Eq. (36). In cases where $\mathbf{Q}(l,k)$ is not positive definite, the nearest positive definite matrix [36] of $\mathbf{Q}(l,k)$ could be used to modify the estimate of $\mathbf{C}_v(l,k)$ used in Eq. (36).

VII. Conclusion and Future Work

In this paper, we proposed three maximum-likelihood-based DoA estimators for a hearing aid system (HAS) which has access to the noise-free target signal via a wireless microphone. The proposed DoA estimators are based on three different models of the direction-dependent relative transfer functions (RTFs) between the HAS' microphones. These RTF models, which we call i) the free-field-far-field model, ii) the spherical-head model, and iii) the measured-RTF model, represent, with increasing accuracy and complexity, the head shadowing effect of the user's head on impinging signals. We showed that the considered signal model and the RTF models allowed the

likelihood function to be calculated efficiently via inverse discrete Fourier transform techniques. In simulation experiments, we analyzed the influences of the true DoA, SNR, window length and reverberation on the performance of the proposed estimators. Moreover, we compared the performance of the estimators with the methods proposed in [18] and [17], which we refer to as the cross-correlation-based method and MLSSL, respectively. The cross-correlation-based method does not take ambient noise characteristics and head shadowing effects into account while MLSSL does take noise characteristics and detailed head shadowing effects into account via a user-specific HRTF database. Simulation results showed that all the DoA estimators proposed in this paper markedly outperform the cross-correlation-based method, while MLSSL outperform the proposed DoA estimators, when the user-specific HRTFs corresponding to the actual location of the target is in the HRTF database used by MLSSL; this is obviously a highly ideal case. We showed that MLSSL is very sensitive to mismatches between the HRTF database and the actual target source distance and the particular user. These mismatches deteriorate the MLSSL performance dramatically while the proposed estimators generally perform well.

Among the DoA estimators proposed in this paper, the measured-RTF-based method provides the lowest DoA estimation error robustly across different noise fields, DoAs, SNRs, and window lengths. In situations where the user-specific measured RTFs or the measured RTFs for a head-and-torso simulator (HATS) are not available, the spherical-head-model-based estimator provides a good performance and is robust against changing physical characteristics and, hence, HRTFs of users.

The proposed estimators rely on spatio-spectral signal characteristics, which are assumed fixed across a short (in the range of milliseconds) duration. It is a topic of future research to extend the estimators to take temporal characteristics of the acoustic scene into accounts, e.g. by modeling the relative movement of the user's head and the target source.

REFERENCES

[1] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[2] A. Bayat, M. Farhadi, A. Pourbakht, H. Sadjedi, H. Emamdjomeh, M. Kamali, and G. Mirmomeni, "A comparison of auditory perception in hearing-impaired and normal-hearing listeners: an auditory scene analysis study," *Iranian Red Crescent Medical Journal*, vol. 15, no. 11, 2013.

[3] J. M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 2, Oct 2003, pp. 1228–1233 vol.2.

[4] J. A. Macdonald, "A localization algorithm based on head-related transfer functions," *Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4290–4296, Jun. 2008.

[5] C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[6] J. Kotus, K. Lopatka, and A. Czyzewski, "Detection and localization of selected acoustic events in acoustic field for smart surveillance applications," *Multimedia Tools and Applications*, vol. 68, no. 1, pp. 5–21, 2014.

[7] S. Goetze, T. Rohdenburg, V. Hohmann, B. Kollmeier, and K.-D. Kammeyer, "Direction of arrival estimation based on the dual delay line approach for binaural hearing aid microphone arrays," in *International Symposium on Intelligent Signal Processing and Communication Systems*, Nov 2007, pp. 84–87.

[8] M. Brandstein and D. Ward, *Microphone Arrays: signal processing techniques and applications*. Springer, 2001.

[9] D. Hoang, H. F. Silverman, and Y. Ying, "A real-time SRP-PHAT source location implementation using stochastic region contraction(src) on a large-aperture microphone array," in *Proc. of IEEE ICASSP*, Apr. 2007, pp. I–121–I–124.

[10] R. Schmidt, "A signal subspace approach to multiple emitter location and spectral estimation," Ph.D. dissertation, Stanford University, 1981.

[11] R. Badeau, G. Richard, and B. David, "Fast adaptive esprit algorithm," in *13th IEEE/SP Workshop on Statistical Signal Processing*, July 2005, pp. 289–294.

[12] J. C. Murray, H. Erwin, and S. Wermter, "Robotics sound-source localization and tracking using interaural time difference and cross-correlation," in *Proc. of NeuroBotics Workshop*, 2004, pp. 89–97.

[13] Y. Huang, J. Benesty, and J. Chen, *Springer Handbook of Speech Processing*. Springer Berlin Heidelberg, 2008, ch. Time Delay Estimation and Source Localization, pp. 1043–1064.

[14] F. Keyrouz, "Advanced binaural sound localization in 3-D for humanoid robots," *IEEE Transaction on Instrumentation and Measurement*, vol. 63, no. 9, pp. 2098–2107, Sept 2014.

[15] C. Vina, S. Argentieri, and M. Rébillat, "A spherical cross-channel algorithm for binaural sound localization," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 2921–2926.

[16] M. Zohourian and R. Martin, "Binaural speaker localization and separation based on a joint itd/ild model and head movement tracking," in *Proc. of IEEE ICASSP*, March 2016, pp. 430–434.

[17] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Maximum likelihood approach to informed sound source localization for hearing aid applications," in *Proc. of IEEE ICASSP*, 2015, pp. 439–443.

[18] G. Courtois, P. Marmaroli, M. Lindberg, Y. Oesch, and W. Balande, "Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions," in *Audio Engineering Society Convention 136*, April 2014, p. 9034.

[19] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Informed TDoA-based Direction of Arrival estimation for hearing aid applications," in *IEEE Global Conference on Signal and Information Processing*, 2015, pp. 953–957.

[20] ——, "Informed direction of arrival estimation using a spherical-head model for hearing aid applications," in *Proc. of IEEE ICASSP*, March 2016, pp. 360–364.

[21] J. Jensen, M. S. Pedersen, M. Farmani, and P. Minnaar, "Hearing system," U.S. Patent 20 160 112 811, April 21, 2016.

[22] S. Braun, W. Zhou, and E. A. P. Habets, "Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions," in *Proc. of IEEE WASPAA*, Oct 2015, pp. 1–5.

[23] R. Duda and W. Martens, "Range dependence of the response of a spherical head model," *Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998.

[24] Y. Avargel, "System identification in the Short-Time Fourier transform domain," Ph.D. dissertation, Israel Institute of Technology, 2008.

[25] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul 2001.

[26] D. R. Brillinger, *Time Series: Data Analysis and Theory*. Society for Industrial and Applied Mathematics (SIAM), 2001.

[27] R. Woodworth, *Experimental Psychology*. Holt, New York, 1938.

[28] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, 2010.

[29] C. I. Cheng and G. H. Wakefield, "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space," in *Audio Engineering Society Convention 107*, September 1999.

[30] R. L. Bouquin-Jeannes, A. A. Azirani, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 484–487, Sep 1997.

[31] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 1–10, 2009.

[32] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997.

[33] P. Kabal, "TSP speech database," Department of Electrical and Computer Engineering, McGill University, Tech. Rep., 2002. [Online]. Available: http://www-mmsp.ece.mcgill.ca/documents/Downloads/TSPspeech/TSPspeech.pdf
[34] D. Avitzour, "Time delay estimation at high signal-to-noise ratio," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 27, no. 2, pp. 234–237, Mar 1991.
[35] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "On the influence of microphone array geometry on hrtf-based sound source localization," in *Proc. of IEEE ICASSP*, April 2015, pp. 439 – 443.
[36] N. J. Higham, "Computing the nearest correlation matrix—a problem from finance," *IMA Journal of Numerical Analysis*, vol. 22, no. 3, pp. 329–343, 2002.

Mojtaba Farmani received the B.Sc. and M.Sc. degrees in Electrical and Computer Engineering from University of Tehran, Iran, in 2009 and 2012 respectively. He is currently pursuing his Ph.D. degree in Electrical Engineering at University of Aalborg, Denmark. He was a Research Assistant at Technical University of Eindhoven, The Netherlands, and also a Visiting Researcher at Delft University of Technology, The Netherlands, and University of Rostock, Germany. His research interests include localization, tracking, statistical signal processing, and audio and speech processing.

Michael Syskind Pedersen received the M.Sc. degree in 2003 from the Technical University of Denmark (DTU). In 2006 he obtained his Ph.D. degree from the department of Informatics and Mathematical Modelling (IMM) at DTU. In 2005 he was a Visiting Scholar at the Department of Computer Science and Engineering at The Ohio State University. Michael's main areas of research are blind source separation and acoustic signal processing including hearing aid signal processing, multi-microphone audio processing and noise reduction. Currently, Michael is a Lead Developer at Oticon A/S, Copenhagen, Denmark, where he has been employed since 2001.

Zheng-Hua Tan (M'00SM'06) received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1999. He is an Associate Professor in the Department of Electronic Systems at Aalborg University, Aalborg, Denmark. He is also a co-founder of the Centre for Acoustic Signal Processing Research (CASPR) at Aalborg University. He was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA, an Associate Professor in the Department of Electronic Engineering at Shanghai Jiao Tong University, and a postdoctoral fellow in the Department of Computer Science at Korea Advanced Institute of Science and Technology, Daejeon, Korea. His research interests include speech and speaker recognition, noise-robust speech processing, multimedia signal and information processing, human-robot interaction, and machine learning. He has authored or co-authored more than 150 publications in refereed journals and conference proceedings. He has served as an Editorial Board Member/Associate Editor for Elsevier Computer Speech and Language, Elsevier Digital Signal Processing, and Elsevier Computers and Electrical Engineering. He was a Lead Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He has served as a Chair, Program Co-chair, Area and Session Chair, and Tutorial Speaker of many international conferences.

Jesper Jensen received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2000, he was with the Center for Person Kommunikation (CPK), Aalborg University, as a Ph.D. student and Assistant Research Professor. From 2000 to 2007, he was a Post-Doctoral Researcher and Assistant Professor with Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. Currently, he is a Senior Researcher with Oticon A/S, Copenhagen, Denmark, where his main responsibility is scouting and development of new signal processing concepts for hearing aid applications. He is a Professor with the Section for Information Processing (SIP), Department of Electronic Systems, at Aalborg University. He is also a co-founder of the Centre for Acoustic Signal Processing Research (CASPR) at Aalborg University. His main interests are in the area of acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.